Final Project Proposal

Scripting for Data Analysis

Due: Friday by 11:59pm ET at the end of Week 6

Planning for the Project

For this assignment, you are to make an initial plan for a project. In the final project you will demonstrate your ability to write Python scripts to access and amass data from fields in one or more of the three types of data studied in the course and to prepare and use data to produce data summaries, lists, and other structures.

1. Choose whether to work individually or to work in a team of two or three people. If you wish to work in a team, specify the people that you have talked with to form a team.

2. Pick a topic of investigation and the data that you will use, ideally from more than one source. The topic could focus on one main data set but also have supporting data. Your topic may focus on a single target topic or person, combinations of them, a comparison of more than one target topic and person, or comparisons over time.

The data may come from any source: those that you have found online, collected from social media, or obtained through other means.

3. Pick several possible methods of analysis in order to give some initial idea of what analysis you will try. This analysis will be to answer the types of questions that you have worked on for the homework assignments. Since we are not focused on visualization, the results of your analysis can be reported as structured tables with a unit of analysis and collected, summarized, or computed values for those units.

The scale of the final project must be larger in scope than the homework assignments in at least one of the following dimensions:

- Incorporating multiple data sets, possibly combining structured, semistructured, or text data
- Conducting additional related analysis questions (either more complex questions or more questions)
- Including additional types of analysis or collecting data, e.g., using another API or social network analysis

4. If you know of places where you may need help with development, try to list that now.

This can range from big things (I want to get information from FourSquare comments) to small things (I'd like a program that helps me to get dates from the documents in my collection and be able to compare dates).

[Potential remaining topics: Social network analysis, geographic locations and maps, getting all Facebook comments. You may request to add to this list.]

5. State in what way you intend for your project to be larger in scope than either of the homework assignments.

6. Based on your plans, you may want to start collecting data.

Assignment Result

Hand in a short document with your initial project plan describing your team, your topic, and your potential methods of analysis, including an assessment of the scope of the project. This document should be a Memo (see templates in Word).

Ideas for Data or Projects

Many websites where people have done analysis also give the sources of their information. For example:

- Nate Silver's 538 website has many examples of analysis. One is this article by Rob Arthur and Jeff Asher on gun violence in Chicago. They say that they got their crime data from the City of Chicago open data portal https://data.cityofchicago.org/.
- Data journalist Yue Qiu has a website with several projects reporting data on workers, trains carrying crude oil, and other statistics from various government websites.

Examples of data sets used by students for the first homework:

- Baseball hall of fame data
- Airbnb data from Kaggle
- Used car test data from the EPA website
- Somerville surveys for sense of safety
- Victim crime data from the Bureau of Justice Statistics
- Data sets from the UPI website: faculty use of Wikipedia data, forest fires, red wine quality, bike rentals

Comparing social media content with real-world events or other items. Examples:

- "Tweet the Debates": Shamma et al. collected tweets associated with political debates and reported on tweet volume over time and social networks of people tweeting.
- "Information Flows in Events of Political Unrest": Nahon and Hemsley compared tweet volume over time with the blogosphere and news stories of events.

- “Toward Predicting Popularity of Social Marketing Messages”: Yu et al. selected restaurants with the most Facebook fans in different categories and analyzed popularity of the posts based on the number of likes and then analyzed the different types of posts by significant (most frequent) words.

- Comparison of different rock bands: Authors collected tweets and Facebook posts over time, looking at user timelines, retweets, likes, number of comments, number of entities, and most frequent words from the text.

- Analysis of tweets around an event: Collect tweets from event hashtags, showing significant words over time, network of people tweeting, and user locations on a map.

Examples of student Final Projects (not all questions are reported, so these do not necessarily reflect the full scope of the student work, and some are multistudent):

Analysis of COVID data: analysis of 2020 rising numbers of covid, analysis 2021 vaccination data.

Questions: What were the numbers for various countries, states, geographical areas? How do the number of vaccinations compare across states, countries, etc?

Data sources: EPA car review data, Edmund’s Car Reviews and Dealership Reviews.

Questions: What are the car dealerships located in the vicinity of Syracuse, and how far away are they? How would you rate American-made automobiles according to mileage, horsepower, fuel efficiency, and cost?

Data sources: Twitter collection from Dave Matthews Band and Phish, including the user profiles and the last 2,000 tweets from their user timelines.

Question: Compare the popularity of the two bands by comparing follower and favorite counts from each profile, average numbers of retweets, and retweets and favorites per followers.

Data source: Tweets collected April 18, 2016, around #parisattacks OR #bataclan with 32K tweets.

Questions: What are the demographics of the tweets? Who are the most influential users (using SNA and retweets)? What are the demographics of the information (looking at the URLs)? What are the sentiments expressed in the tweets?

Data source: Tweets about Boston Red Sox and NY Yankees, and Facebook posts and comments from the two teams’ Facebook fan pages.

Questions: Updated analysis of many questions based on earlier article by Bialik in Five-thirtyeight for 2014.

Data sources: Airbnb data set, collected tweets about Airbnb.

Questions: What factors influence the customer review scores? How much money can each host make in a particular time period? Can we use tweets about airbnb to discover recent popular travel trends?

Data sources: MovieLens data set with 100K reviews and selected movie reviews downloaded from IMDb in HTML.

Questions: Do movie ratings differ according to gender and genre? Do movie reviews differ by gender for movies with male or female protagonists?

Data sources: Tweets around the topic #techno, user profiles from SoundCloud in the genre techno, from the API.

Question: Compare the demographics of the Twitter users and the SoundCloud users.